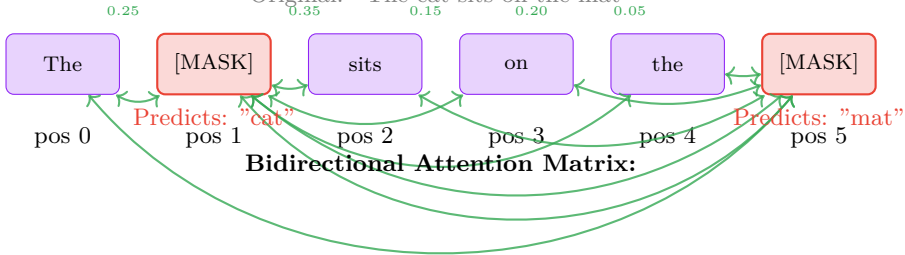


Bidirectional Attention Patterns with MASK Tokens

MLM Training Example:

Original: "The cat sits on the mat"



Bidirectional Attention Matrix:

	The	[MASK]	sits	on	the	[MASK]
The						
[MASK]						

Key Differences from Autoregressive Models

Bidirectional Context: No Causal Masking: Parallel Prediction:

Uses both left & right context Future tokens inform predictions Multiple masks predicted jointly

MLM Objective: Predict original tokens using full bidirectional context

